Multilevel autoregressive models when the number of time points is small

Fien Gistelinck

Department of Data Analysis, Ghent University, Belgium

Tom Loeys

Department of Data Analysis, Ghent University, Belgium

Nele Flamant

Department of developmental, personality and social psychology, Ghent University, Belgium

Author Note

This research was supported by a grant of the Ghent University Special Research Fund, grant number BOF.STA.2015.0008.01.

Abstract

The multilevel autoregressive model disentangles unobserved heterogeneity from state-dependence. Statistically, the random intercept accounts for the dependence of all measurements on an observed underlying factor, while the lagged dependent predictor allows the value of the outcome to depend on the outcome at the previous time point. In this paper we consider different implementations of the simplest multilevel autoregressive model, and explore how each of them deal with the endogeneity assumption and the initial conditions problem. We discuss the performance of the no centering approach, the manifest centering approach and the latent centering approach in the setting where the number of time points is small. We find that some commonly used approaches show bias for the autoregressive parameter. When the the outcome at the first time point is considered predetermined, the no centering approach assuming endogeneity performs best.

Keywords: Structural Equation Modeling, Latent Centering, Multilevel Autoregressive Models, Panel Data

Multilevel autoregressive models when the number of time points is small

Introduction

Longitudinal data modeling often focuses on modeling a trend over time. Latent growth models for example are useful to study developmental processes, but may be less useful when studying daily or weekly affective measurements. Indeed, for the latter a trend over time is likely to be absent. Still, the value of an outcome (e.g., positive mood) at a particular time point may depend on its value at a previous time point. This can be viewed as state-dependence (Figure 1a): positive mood on a particular time point has an effect on positive mood at the next time point. However, there may also be some unobserved underlying trait that entirely explains the correlation between subsequent measures (Figure 1b), for example personality affecting the repeated measures of mood. Or, more realistically, the true underlying process might be a combination of state-dependence and trait (Figure 1c).

The multilevel autoregressive model that we will discuss in this paper nicely combines both the state-dependence and trait. From a statistical perspective, the lagged dependent variable in this model captures the state-dependence while the random intercept represents the underlying trait that affects all measurements equally. In this paper, we will consider a simple version of this model, but more complex variations are possible as well. First, we will only allow the immediate prior value to influence the current value (i.e., an autoregressive process of order 1 or AR(1) process is assumed), while one could permit earlier lagged values to affect the current value as well. Second, we will not consider effects of time on the outcome of interest either (e.g. no latent slopes as in autoregressive latent trajectory models). Third, we assume equally spaced measurements and time-constant autoregressive effects that are homogeneous between subjects. Fourth, we will limit our attention to a single outcome variable and do not include any other time-constant or time-varying predictors in the model. We make all those simplifying assumptions to clearly focus on two well-known issues in the estimation of the multilevel autoregressive model that occur even in this simplest model: the initial conditions problem and the endogeneity problem. In the

next paragraphs we elaborate on both issues in more detail.

Since the multilevel autoregressive model focuses on modeling processes that are ongoing, the measurement at the initial time point is affected by an unavailable presample response. This is commonly referred to as the 'initial conditions problem'. Usually the response at the initial time point is treated as predetermined (Bollen & Curran, 2004). Alternatively, one can interpret the initial conditions as a missing data problem within the Bayesian framework (Asparouhov, Hamaker, & Muthén, 2018) and perform imputations for the missing presample response. Following Zhang and Nesselroade (2008) one can treat for example the missing presample responses as auxiliary parameters that have their own data-dependent priors.

One of the basic assumptions of regression analysis is the exogeneity assumption stating that residuals are independent from the predictors in the model. However, in the multilevel autoregressive model this assumption is violated, and results in what is known as 'endogeneity'. Indeed, because the random intercept is a direct cause of the outcome at every point in time, there is an intrinsic correlation between the lagged dependent variable and the residual (i.e., the combination of the random intercept and lower level error term in this case) at every time point. This is especially important when the outcome at the first time point is only considered as a predictor. When the endogeneity is not appropriately dealt with, this may lead to bias in the estimation of the model parameters. The result of endogeneity tends to be that the effect of the lagged dependent variable is estimated as too large (Allison, Williams, & Moral-Benito, 2017). A fixed effects (or group-mean centering) approach can be used in some settings to tackle endogeneity (Hamaker & Muthén, 2019; Loeys, Josephy, & Dewitte, 2018), but fails in models with a lagged dependent variable as predictor and is known to be prone to Nickell's bias (Nickell, 1981).

The multilevel autoregressive model originally stems from the time series literature in economics, but the last couple of years it has received a lot of attention in the psychology literature as well. Especially the development of dynamic structural equation modeling (DSEM) in Mplus has enabled researchers to address longitudinal

research questions of increasing complexity. The focus of most DSEM-applications is on intensive longitudinal data. The richness of such long time series data allows to consider more complex multilevel autoregressive models with, among others, subject-specific autoregressive parameters, subject-specific error variances,... . All those features are nicely present in DSEM in Mplus. At first sight, there are no real barriers to fit the simple multilevel autoregressive model that we consider in this paper with DSEM. However, in this paper here, we want to focus especially on the setting where the number of time points is small and investigate the two aforementioned issues in more detail. For example, as DSEM treats the initial conditions problem as a missing data problem, we wonder how small the number of time points may be for DSEM to still perform well. Asparouhov et al. (2018) acknowledge this potential issue: 'Note that when the time series model is sufficiently large with 30 or more observations, it is very unlikely that the prior specification affects the estimation. The effect of this prior tends to fade away beyond the first few time periods. However, when the number of time periods in the time series is small, such as less than 20, one can expect that the prior will have some small effect on the estimates. The burn-in phase prior estimation method we propose here appears to be working quite well even for short time series.' However, these authors do not study the impact in detail. Second, the authors claim that 'the DSEM latent centering approach resolves Nickell's bias and that the latent centering is superior to the observed centering'. Given that the initial conditions problem and endogeneity are intertwined, it is not so clear however how the imputation of the missing presample response and latent centering affect each other.

Since the consequences of violation of the exogeneity assumption and the impact of imputing missing pre-sample response as a way to overcome the initial conditions problem will become most apparrent when the number of time points is small, we will focus here in this paper on that particular setting. Although DSEM might primarily have been developed for more complex models with intensive longitudinal data in mind, studying its performance when the number of time points is small and contrasting its performance with other estimation approaches, increases knowledge about its

properties. Furthermore, while the simple version of the multilevel autoregressive model that we consider here may be of limited practical relevance, it is important to note that this simplest model can be viewed as half of a random intercept cross-lagged panel model (RI-CLPM) (Hamaker, Kuiper, & Grasman, 2015) without cross-lagged effects, or as an autoregressive latent trajectory model (ALT) (Bollen & Curran, 2004) without a linear effect of time. The RI-CLPM and ALT model are very often applied to panel data, which typically have a rather large number of individuals and a rather small number of time points. It is exactly in these circumstances that the aforementioned issues may be most present. Hence, our findings on the performance of the different estimation strategies for the simple multilevel autoregressive model when the number of time points is small will have important implications for the estimation of more complex RI-CLPM and ALT-models as well.

The multilevel autoregressive model

As a motivating example throughout this paper we consider the study of Flamant and Soenens (n.d.). These authors investigate pupil's perception of controlling behavior of the teacher and how it affects their motivation, class engagement, psychological need-frustration and general well-being. About 400 pupils from different schools and grades were asked to fill in a weekly questionnaire for four consecutive weeks. Here, we focus on the estimation of the dynamics of the autonomous motivation of the pupil with only those four repeated measures available, and explore to what extent the motivation of last week affects the current motivation. Before we start describing the model in more detail, it is important to note that for identification reasons this minimum of four waves is required for the multilevel autoregressive model under the assumption that the autoregressive parameter is constant over time and a stationary process (Bollen & Curran, 2004), regardless of the modeling framework that is used.

We denote by y_{ti} the motivation of pupil i (level 2) at time t (level 1). As argued in the introduction the true underlying process describing the motivation of the pupil may be a combination of state-dependence and trait. We therefore start by

decomposing y_{ti} into an individual-specific contribution μ_i (the trait or equilibrium) and the deviation of individual i at time t, denoted z_{ti},

$$y_{ti} = \mu_i + z_{ti} \tag{1}$$

with i referring to the subject number ($i = 1, \ldots, N$) and t to the time point ($t = 1, \ldots, T$). We can further describe the individual's equilibrium as a grand mean with white noise:

$$\mu_i = \mu + \eta_i \tag{2}$$

with $\eta_i \sim N(0, \tau_\eta^2)$. To capture the state-dependence, the individual's deviations at level 1 can be modeled as a first-order autoregressive structure with autoregressive parameter ρ ($-1 < \rho < 1$):

$$z_{ti} = \rho z_{(t-1),i} + \varepsilon_{ti} \tag{3}$$

with $\varepsilon_{ti} \sim N(0, \sigma_\varepsilon^2)$. As the residual ε_{ti} accounts for the part that is not predicted by the previous $z_{(t-1),i}$, it is also referred to as *the innovation*. For instance, in the study of Flamant and Soenens (n.d.), a large residual variance σ_ε^2 would correspond to a high level of perturbation in the autonomous motivation of the participants. While an autoregressive parameter ρ close to zero would imply little carryover effect of motivation from one measurement occasion to the next, an autoregressive parameter close to one would imply that there is considerable carryover between consecutive measurement occasions. Hence, the autoregressive parameter is often referred to as a measure of *inertia*. If the stability of constructs shows some time-invariant characteristic (e.g., due to the intrinsic motivation of the student), it is shown that the autoregressive parameter alone is not able to account for this trait (Hamaker et al., 2015). Even in case the autoregressive parameter is very close to one, indicating a very high carryover effect, the parameter can only account for temporal stability as its effect nullifies when enough time passes.

As already mentioned in the introduction we make several simplifying assumptions. For example, model (3) assumes the autoregressive parameter ρ and the level 1 residual variance σ_ε^2 to be fixed instead of random. As mentioned by Jongerling,

Laurenceau, and Hamaker (2015), a person-specific residual variance might be important as it captures differences in sensitivity and/or exposure to innovation. That is, some individuals may show more variability than other. Assuming the residual variance as fixed when in reality it is individual-specific may impact the estimation of the autoregressive parameter as well. This is explained by the fact that the total variance in a ML-AR(1) model both depends on the innovation variance and the autoregressive parameter, and hence wrong assumptions about the former may bias the estimation of the latter. However, since we are focusing on panel data with a small amount of time points however, not enough information may be available in the data to reliably estimate a subject-specific variance, and therefore we will treat the residual variance as fixed. For similar reasons, we also assume a fixed rather than a random autoregressive parameter. As shown by Schultzberg and Muthén (2018), a random autoregression parameter requires a lot more time points in order to obtain a model fit with good performance. Incorrectly treating the autoregressive parameter as fixed rather than random introduces no bias in the mean structure parameter estimates (Baird & Maxwell, 2016), but distorts the standard errors. Finally, we consider a stationary process, meaning that the mean and variance of the outcome variable are stable over time for each individual. If the AR(1) process is stationary, the autoregression parameter ρ will lie within a range of -1 to 1. In this paper we will not focus on the consequences of violations of the above simplifying assumptions but we investigate how different implementations to fit the simplest multilevel autoregressive model deal with the endogeneity and initial conditions issue.

Implementations of the ML-AR(1) model

In this section the most frequently used approaches for fitting the ML-AR(1) model will be introduced. Upon noting that $z_{(t-1),i}$ can be rewritten as $y_{(t-1),i} - \mu_i$ and substituting (3) into (1), we find that

$$y_{ti} = \mu_i + \rho(y_{(t-1),i} - \mu_i) + \varepsilon_{ti} \tag{4}$$

or equivalently

$$y_{ti} = \alpha_i + \rho y_{(t-1),i} + \varepsilon_{ti} \tag{5}$$

with $\alpha_i \sim N(\alpha, \tau_\alpha^2)$, $\alpha = (1-\rho)\mu$ and $\tau_\alpha^2 = (1-\rho)^2\tau_\eta^2$.

Separating the observed y_{ti} into an unobserved between-subject component and within-subject component as in model (1), and specifying models as (2) and (3) for these components, is the approach that is taken in the multilevel SEM framework. Explicitly modeling the observed y_{ti} as in model (5) on the other hand is the approach taken in the 'traditional' multilevel framework. As we have just shown, there is a clear one-to-one relationship between both approaches for the simple multilevel autoregressive model that we consider here. It is however important to note that the intercept α_i in model (5) does not reflect the subject-specific equilibrium but rather a transformation of it.

The traditional multilevel framework

No centering. Considering model (5) it may be tempting to use the traditional multilevel framework and naively fit this model using the *lmer*-function from the *lme4*-package in *R* for example,

```
fitNCEXO <- lmer(yy ~ 1 + yylag1 + (1|id) , data = datlong)
```

or with the *lme*-function from the *nlme*-package in *R*:

```
fitNCEXO <- lme(yy ~ 1 + yylag1, random=~1|id, data = datlong)
```

with data in long format (*datlong*), and *yy* containing the observed outcomes y_{ti}, $yylag1$ the lagged outcomes and *id* the subject identification. Obviously, other standard multilevel modeling software, such as HLM (Raudenbush, 2004) or SAS's proc mixed (SAS Institute, 2008) could have been used instead. Most importantly, with regard to the initial conditions problem all those implementations of the standard multilevel approach treat the lagged dependent variable at time point 1 as missing (i.e y_{0i} is missing), and hence the outcome at time point 1 (i.e. y_{1i}) is considered predetermined. Furthermore, the standard multilevel approach assumes exogeneity, that is the random

intercept is independent from the predictors in the model. This is especially problematic for the lagged predictor y_{1i} as it is treated as predetermined and independent from the random intercept. These assumptions are made more explicit in Figure 2: the observed outcomes Y_1 till Y_4 are presented in rectangular boxes, while the unobserved random intercept is presented by an oval form. No arrow is drawn from the random intercept to Y_1 (i.e. the exogeneity assumption is made) and Y_1 is treated as predetermined. We will refer to this approach as 'No centering - Exogeneity' (abbreviated NC-EXO).

Manifest centering. To solve the endogeneity problem, scholars have proposed to use the fixed effects approach rather than the random effects approach for the intercept. In fact, this amounts to eliminating the effect of the time-constant subject-specific trait. This can be achieved by cluster-mean centering, also referred to as manifest centering or observed centering. This is a fruitful strategy (Hamaker & Muthén, 2019; Loeys et al., 2018) when a predictor and an outcome share some unmeasured time-constant subject-specific cause (or upper-level confounder), but as explained below this is problematic with a lagged dependent variable as predictor.

Jongerling et al. (2015) suggested to use the observed individual's sample mean $\bar{y}_i$ to center the lagged outcome variable in (5):

$$y_{ti} = \mu_i + \rho(y_{(t-1),i} - \bar{y}_i) + \varepsilon_{ti} \tag{6}$$

This model can also easily be fitted with the *nlme*-package for example:

```
fitCMC <- lme(yy ~ 1 + yylag1c, random=~1|id, data = datlong)
```

where now $yylag1c$ is the cluster-mean centered lagged dependent variable. A graphical representation of the model can be found in Figure 3. The cluster-mean centered lagged predictors Y_c^1, Y_c^2 and Y_c^3 are represented by rectangular boxes. It is immediately clear that also in this approach the first outcome variable is treated as predetermined. We will refer to this approach as cluster mean centering, abbreviated CMC. An alternative to the observed cluster mean centering is based on a two-step procedure, in which an estimate for the random intercept is first obtained by fitting an empty model. Here, we will no further discuss this alternative as Jongerling et al. (2015) showed that the

differences in performance between both approaches are negligible. More importantly, Nickell (Nickell, 1981) showed that the CMC-approach suffers from bias for the autoregressive parameter ρ. An approximation of this bias is given by the following formula:

$$-\frac{1+\rho}{T-1} \tag{7}$$

The bias will be especially present when the number of time points T is small. Given that expression (7) does not depend on N, it is obvious that Nickell's bias is persistent in case the number of subjects increases. Interestingly, simulation studies by Hamaker et al. (2015) also showed that, even in case the true underlying mean in model (6) is used to center, bias for the autoregressive parameter is still observed.

The Structural Equation Modeling framework

No centering. We noted above that in the traditional multilevel framework the exogeneity assumption is made by default. This limitation can however easily be overcome in the SEM-framework. By viewing the random intercept in the multilevel framework as a latent variable in the SEM-framework, we can specify equivalent models. More importantly, the SEM-framework allows to specify a correlation between the outcome at the first time point and the random intercept, hereby solving the endogeneity issue (Allison et al., 2017). This approach, which we will refer to as 'No Centering - Endogeneity', abbreviated NC-ENDO, is visualized in Figure 4. In the absence of the correlation between the outcome at the first time point and the random intercept, the approach reduces to NC-EXO. Also in this approach the outcome at the first time point is predetermined and does not depend on an unobserved presample response.

This approach can easily be implemented with SEM-software such as EQS (Bentler, 2004), LISREL (Jöreskog & Sörbom, 1996), OpenMx (Boker et al., 2011), Mplus (Muthén & Muthén, 2012), Stata's gllamm (Rabe-Hesketh, Skrondal, & Pickles, 2004), SAS's proc CALIS (SAS Institute, 2013) or R's `lavaan` (Rosseel, 2012). We give some example code for the latter here. Assuming that data are organized in a wide data

format now rather than long format (with $yy.1$ till $yy.4$ denoting the outcomes at time 1 to 4), we can implement the model with 4 time points shown in Figure 4 as follows

```
modelNCENDO <- '
# random effects
ri =~ 1* yy.2 + 1* yy.3 + 1* yy.4
ri ~~ vari *ri + delta *yy.1
# mean structure
yy.1 ~ alpha0 *1
yy.2 ~ alpha *1 + rho *yy.1
yy.3 ~ alpha *1 + rho *yy.2
yy.4 ~ alpha *1 + rho *yy.3
# residual covariance model
yy.1 ~~ resvar0 *yy.1
yy.2 ~~ resvar *yy.2
yy.3 ~~ resvar *yy.3
yy.4 ~~ resvar *yy.4 '
fitNCENDO <- sem(modelNCENDO, data = dat_wide )
```

While the endogeneity issue is addressed in this approach (by specifying the covariance *delta* between the random intercept ri and the first measurement $yy.1$ in the above code), it should be noted that in the above code the outcome at the first time point is still treated as predetermined, and has its own mean and variance (denoted by *alpha0* and *resvar0*, respectively).

Latent centering. As explained above, the usual within-between decomposition (μ_i and z_{it} in model (1)) is fundamental to two-level structural equation modeling. The within-component z_{it} can be viewed as a centered variable since $z_{it} = y_{it} - \mu_i$. Rather than using the observed mean to center the variables, the latent mean can indeed be used to avoid bias due to unaccounted sampling error (Lüdtke, 2008). The latent centering that is inherently present in multilevel SEM is also immediately clear from model (4). The other important question relates to the treatment of y_{1i} in multilevel

SEM software such as DSEM. When treated similarly as all outcomes y_{ti} at other time points (t=2,. . . ,T), y_{1i} depends on unobserved y_{0i} (as in model (4)). To solve the initial conditions issue, DSEM treats z_{i0}, the within-part of y_{i0}, as an auxiliary parameter that has its own prior within the Bayesian framework, More precisely and following Zhang and Nesselroade (2007), DSEM estimates the prior during a burn-in phase of the Markov Chain Monte Carlo (MCMC) estimation. In the first iteration z_{0i} is set to zero, but after each MCMC iteration during the burn-in phase of the estimation a new prior is computed as the normal prior with mean and variance the sample mean and variance of z_{ti} over all $t > 0$. Hence, in contrast to the 3 other approaches outlined above, DSEM does not treat the measurement at the first time point as predetermined. We will refer to this approach as the DSEM or *LC-ENDO* approach. The corresponding model is visualized in Figure 5: every observed variable is now decomposed into a within-part and a between-part, represented by the rounded boxes at the within (below the dashed line) and the between (above the dashed line) level. The unobserved pre-sample response at the within-level is represented by a circle, and affects the measurement at the first time point. The Mplus DSEM-code to fit model (4) can be specified as follows

```
TITLE: LC_model
DATA: FILE = "datlong.dat";
VARIABLE:
 NAMES = id yy;
 MISSING=.;
 cluster = id;
 lagged = yy(1);
ANALYSIS:
 TYPE IS TWOLEVEL;
 estimator = Bayes;
MODEL:
 %WITHIN%
  yy (a);
```

```
  yy on yy&1 (b);
 %BETWEEN%
  [yy] (c);
  yy (d);
```

As already mentioned before, in contrast to the 3 approaches discussed above the DSEM approach relies on the Bayesian framework rather than the maximum likelihood estimation framework. In principle the *NC-EXO, NC-ENDO* and *CMC*-approach can be implemented in a Bayesian framework too, but this will not be considered here.

Empirical Example

To illustrate the four above described approaches, we applied them to our motivating example data on autonomous motivation of primary school pupils. For ease of comparison, we eliminated the incomplete cases, so that 310 pupils with complete data remained. We used the following statistical frameworks and software to implement the four approaches:

$$\begin{aligned} &\textit{NC-EXO} \text{ within the MLM framework using the } \texttt{nlme}\text{-package in R,} \\ &\textit{CMC} \text{ within the MLM framework using the } \texttt{nlme}\text{-package in R,} \\ &\textit{NC-ENDO} \text{ within the SEM framework using the } \texttt{lavaan}\text{-package in R,} \\ &\textit{LC-ENDO} \text{ within the multilevel SEM framework using DSEM in Mplus} \end{aligned} \tag{8}$$

The model within Mplus was fitted using the R-package `MplusAutomation` (Hallquist & Wiley, 2018). In the Bayesian approach, prior distributions, number of burn-in iterations, etc. need to be specified. All specifications were kept at default. The code for the models can be found in the supplementary materials.

The results in Table 1 reveal remarkable differences between the four approaches. The *NC-EXO* and *LC-ENDO*-approach both find a very strong autoregressive effect, (larger than 0.70) while the estimated autoregressive parameter is close to 0.30 for the *NC-ENDO* approach. The estimated negative autoregressive effect by the *CMC* approach is not realistic and may be indicative for the earlier mentioned Nickell's bias.

Also, the residual variances are comparable between the *NC-EXO* and *LC-ENDO*, smaller for the *NC-ENDO*, and even more so for the *CMC model*. The intercepts and random intercept variances are only directly comparable within the no centering and centering approaches, respectively. However, after transforming the intercept α into μ based on $\alpha = (1-\rho)\mu$, all approaches estimate the equilibrium μ for autonomous motivation around 3.35. The *NC-EXO model* finds no evidence for a positive random intercept variance (i.e., estimate at the boundary of the parameter space), while also in the *LC* approach, the estimated random intercept variance is small. The estimate for the random intercept variance under the *NC-ENDO model* equals $\frac{0.34}{(1-0.29)^2} = 0.67$, implying that 95% of the participants specific equilibrium lie between 2.09 and 4.66.

Simulation Study

Given the discrepancies between the four approaches in our empirical example, we explore their performance in a simulation study where the underlying truth is known. We generate the y_{ti}'s from a multivariate normal distribution as follows:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{Ti} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_i \\ \mu_i \\ \vdots \\ \mu_i \end{pmatrix}, \sigma_Y^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{pmatrix}\right) \tag{9}$$

with $\mu_i \sim N(\mu, \tau_\mu^2)$. The conditional distribution of y_{ti} given $y_{(t-1),i}$ then equals $N\left(\mu_i + \rho(y_{(t-1),i} - \mu_i), (1-\rho^2)\sigma_Y^2\right)$, or equivalent $N\left((1-\rho)\mu_i + \rho y_{(t-1),i}, (1-\rho^2)\sigma_Y^2\right)$ for $t = 2, \dots, T$. The parameters ρ, μ, σ_Y^2 and τ_μ^2 were set to 0.4, 2, 4 and 3, respectively. Those choices imply that the true parameter values for σ_ϵ^2 and τ_η^2 are $(1-0.4^2) \times 4 = 3.36$ and 3 in model (4); and the true parameter values for for α and τ_α^2 are $(1-0.4) \times 2 = 1.2$ and $(1-0.4)^2 \times 3 = 1.08$ in model (5), respectively. The number of subjects was fixed at 50 (we also ran simulations with N=200 and reached similar conclusions) and the number of time points was varied from 4 till 20 by steps of 2. We compare the four approaches (NC-EXO, MC, NC-ENDO and LC-ENDO) using 200 replications for each setting.

Estimation with DSEM for the LC-ENDO approach is performed in a Bayesian framework, which requires specification of prior distributions. We choose to use the default priors of DSEM in Mplus:

$$\begin{aligned} \mu &\sim N(0, \infty) \\ \rho &\sim N(0, \infty) \\ \tau_\eta^2 &\sim IG(-1, 0) \\ \sigma_\varepsilon^2 &\sim IG(-1, 0) \end{aligned} \tag{10}$$

with IG referring to the inverse gamma distribution. Mplus thus allows for improper distribution as they are advantageous in small samples with respect to bias and mean squared error (Asparouhov & Muthén, 2010). The $IG(-1, 0)$ is approximately equivalent to the uniform distribution with minimum 0 and maximum ∞.

When considering four time points, DSEM reported convergence issues for 29 of the 200 repetitions. Due to non-convergence, unusually high parameter estimates were sometimes found, if a value was reported at all. As a result, we opted to limit the sensitivity of these outliers by using robust summary statistics. The median parameter estimates can be found in Figure 6, and a similar plot for the median absolute errors (MedAE) can be found in the supplementary materials. As one can see, all approaches, except for the *CMC* model, tend to yield unbiased estimates for the model parameters when at least 10 time points are available. This is consistent with the findings of the simulation study of Hamaker and Grasman (2015). The *CMC* model shows negative bias for the autoregressive parameter, as predicted by Nickell's bias, even when the amount of time points is as high as 20 (the prediction based on formula (7) is indicated by the blue triangle). The *CMC* approach also yields biased estimates for the random intercept variance.

Upon contrasting *NC-EXO* with *NC-ENDO* the consequences of ignoring the endogeneity issue become clear. As the results for *NC-ENDO* reveal, the bias due to violation of the exogeneity assumption in *NC-EXO* can be eliminated by adding a correlation between the first lagged outcome and the random intercept. We also see that as soon as the number of time points is smaller than 6, *NC-EXO* estimates the random

intercept as almost zero, causing maximal bias in all other variables (cf. the hitch at the start of the curves). The results for *NC-ENDO* also illustrate that assuming the first outcome variable as a predetermined variable does not bias the parameters. Using Mplus' DSEM to fit *LC-ENDO*, we notice that DSEM does not perform well when the number of time points is smaller than 10. This is interesting because no bias was found in a similar simulation study in Asparouhov et al. (2018) investigating latent centering in the setting of intensive longitudinal data (i.e. with much larger T). The question arises what causes the bias. Is it possible that latent centering is not suitable when the number of time points is rather small? Does adding the auxiliary parameter for the missing presample response in order to resolve the initial condition problem make the estimation unreliable? Could it be that the bias is attributed to specific choices made in the default settings for the priors in DSEM? For instance, what is the impact of the choice of the priors? In order to elucidate the performance of DSEM, we implemented the *LC-ENDO* approach using an open source Bayesian software package, more specifically, the R-package `Rjags`. The code for the Bayesian models that we discuss in the next section, can be found in the supplementary materials.

Alternative models with latent centering

In this section we try to find out the source of the biased parameter estimates for ρ with DSEM when the number of the time points is small, and consider different implementations of model (4) in `Rjags`. Firstly, the bias might be due to the endogeneity problem, which may not be resolved by latent centering. In a first implementation, the outcome variable is split up into a latent between and within part at all time points. However, the outcome at the first time point does not contribute to the estimation of the between-effect. Hence, the exogeneity assumption is made here. The outcome at the first time point has its own prior distribution based on the observed mean and variance of the y_{it}'s, and is treated as predetermined. Figure 7a represents this *BAY-LC-EXO model*. In a second implementation, we no longer make the exogeneity assumption. In addition, we consider an auxiliary parameter for the

unobserved presample responses, see Figure 7b. The auxiliary parameter is introduced at the within-level as a direct start-up of the autoregressive process, and its prior is based on the mean and variance of the outcomes at other time points. We will refer to this model as the *BAY-LC-ENDO-0 model*. The number '0' at the end of this acronym represents the time point used as the start-up of the autoregressive process.

A second possible explanation for the bias seen with DSEM is that the introduction of this auxiliary parameter may make the estimation unreliable in case of few time points. For this reason, a third model implementation is introduced, which removes the auxiliary parameter from the *BAY-LC-ENDO-0 model*. A graphical representation of this model, which we will refer to as the *BAY-LC-ENDO-1 model*, can be found in Figure 7c. In this case, the residual variance of the within-part of the first outcome variable is considered free. As a result, the endogeneity problem has been dealt with, while the first time point represents the start-up of the autoregressive process, hence the '1' at the end of the acronym.

The issues above are all related to the specification and assumptions of the model. However, it is possible that the origin of the bias lies within the statistical framework used. Bayesian methods do not rely on asymptotics, a property that can be a hindrance when employing frequentist methods in small sample contexts. Although Bayesian methods are better equipped to model data with small sample sizes, estimates are highly sensitive to the specification of the prior distribution (McNeish, 2016). Therefore, non-informative prior distributions are often suggested. The most frequently used non-informative prior distributions for mean structure parameters, like μ and ρ in our case, is the normal distribution with mean zero and a high variance. For the variance structure parameters, such as τ_η^2 and σ_ε^2, several non-informative prior distributions have been proposed, such as an inverse gamma distribution with a small scale parameter, a uniform distribution with a large scale parameter, a half-Cauchy distribution or a log-normal distribution. Alternatively, one may specify a prior distribution for the standard deviations instead, for example, using a uniform distribution (Zitzmann, Lüdke, & Robitzsch, 2015).

While the Bayesian approach is often used in small sample contexts, it can provide less accurate estimates in case of challenging data constellations, such as small clusters, a situation inherent to panel data (Zitzmann, Lüdtke, Robitzsch, & Marsh, 2016). Recently, McNeish (2019) showed in the context of DSEM that the likelihood which updates the prior distribution carries less weight within the posterior distribution in case of small samples. As a result, the non-informative prior distributions may become unintentionally informative (McNeish & Stapleton, 2016). The reasoning can be similar for small cluster sizes. Therefore, we also considered a frequentist approach which does not use prior distributions, but relies on maximizing the likelihood. Within the traditional SEM framework, one can still perform latent centering. This is achieved by defining a latent variable μ_i over all time points and a residual latent variable z_{ti} for each time point (Beaujean, 2014). More specifically, we build on Kenny and Zautra's Trait-State-Error model (also known as the STARTS model; Kenny & Zautra, 2001), which used an SEM approach to decompose a person's measured level on some psychological characteristic at a particular time into a component reflecting their typical level, a component reflecting their true current state, and a component reflecting measurement error. Like Kenny and Zautra (2001), we distinguish here as well stable and time-varying sources of dependence, but we do not make adjustments for measurement error. In practice, one needs to eliminate the original residuals of the outcome variables by (manually) setting their variance to zero such that the model is identified. In the SEM framework, it is then possible to deal with the endogeneity problem as well, by adding a correlation between the latent intercept and the latent within-part of the first outcome variable. Moreover, the variance of its within-part can be considered free, allowing the first time point to represent the start-up of the autoregressive process. Hence, we end up with a fourth model similar to the *BAY-LC-ENDO-1 model*, that can be estimated within the maximum likelihood framework, the *ML-LC-ENDO-1 model*. A graphical representation of this implementation can be found in Figure 7d.

Simulation study 2

Consider the data-generating process of simulation study 1 as described by expression (9). Again, we assume 50 subjects (we also ran simulations with N=200 and reached similar conclusions) and 4 till 20 time points, varied by steps of 2. In this second simulation study, the four models depicted in Figure 7 were fitted with `Rjags` and `lavaan`, and were compared to Mplus' DSEM. As `Rjags` neither allows for infinite bounds for prior distributions nor improper priors, we decided to fix the prior distribution for the mean structure parameters to a normal distribution with variance 100,000 in both `Rjags` and Mplus. We also considered three different prior distributions for the variance structure parameters: the uniform distribution with scale parameter 1,000, the inverse gamma distribution with scale parameter 0.001 and the half-Cauchy distribution with scale parameter 10 (not possible in Mplus). Based on the model diagnostics, such as traceplots, the Brooks-Gelman-Rubin plots and the autocorrelation function (Albert, 2009), we decided to keep the thinning fixed at one, to increase the number of chains from two to three and to compute 50,000 MCMC iterations. Note that `Rjags` allows the user to define the burn-in phase separately from the update phase. In the latter phase, `Rjags` also makes a difference between adapting the samplers used in the Markov chain and the initial burn-in period (Plummer, 2015). This is in contrast to Mplus, which uses by default only half of the number of iterations as a burn-in phase and the other half to compute the posterior distribution. Depending on the distance of the Proportional Scale Reduction (PSR) from one (i.e. the convergence criterion), the posterior distribution might be based on a smaller amount of iterations. In order to maintain as much equivalence between `Rjags` and Mplus, the same amount of iterations for the update phase in both software-packages were imposed, namely 25,000 (half of 50,000), using "FBITER" within Mplus (Muthén, 2010).

From the simulation study comparing the different priors (not shown here), it was clear that the uniform and half-Cauchy distribution provided similar results for all `Rjags` models. The diagnostics of the models based on the inverse gamma prior were bad, which might explain the difference in performance compared to the other priors in

`Rjags` models and Mplus' DSEM. Hence, we will further only focus on the result assuming a uniform prior distribution.

The simulation results for the median of the parameter estimates are depicted in Figure 8. A similar plot for the MedAE can be found in the supplementary materials. As Mplus reported convergence issues for 84 of the 200 repetitions in the setting with four time points, the impact of these runs on the results are limited with these robust measurements. Still, this non-convergence issue may explain the jump of the different implementations when going from four to six time points.

Figure 8 reveals that *BAY-LC-EXO*, which ignores the endogeneity problem, performs rather similar to DSEM. To keep the readability of the figure, the 95%-confidence intervals based on the median absolute differences (MADs) were plotted for only these two models. There are some differences in the autoregressive parameter and when $T = 4$ in the other model parameters, although the trend is very akin to one another. Figure 8 further reveals that *BAY-LC-ENDO-0* attenuates the bias for the estimate of the autoregressive parameter compared to the DSEM and *BAY-LC-EXO* . The estimator of the intercept from *BAY-ENDO-0*, however, shows more bias when the number of time points is rather small. When *BAY-LC-ENDO-1* is used, it performs very well, even in settings with a small number of time points. This confirms the fact that latent centering can be used in order to resolve the endogeneity issue, but that treating the first outcome as predetermined performs better. Moreover, it can be seen that the corresponding ML implementation (*ML-LC-ENDO-1*) provides unbiased estimates as well and might be preferred over its BAY implementation given that is has less convergence issues when the number of time points is very small.

To summarize, we have found two approaches for fitting the multilevel autoregressive model that show no bias for the autoregressive parameter when applied to panel data. On the one hand, if the researcher is not interested in the intercept, an uncentered model in which a correlation between the first lagged outcome variable and the random intercept is included, can be used (i.e., the *NC-ENDO model*). On the other hand, if the researcher is interested in directly estimating the equilibrium, latent

centering could be used. In this case, the *ML-LC-ENDO-1 model* based on maximum likelihood in the SEM framework performs best.

Discussion

In this paper, we discussed different ways to fit the ML-AR(1) model in light of the endogeneity problem and the initial conditions problem. We found that ignoring the endogeneity problem can lead to severe bias in the autoregressive parameter. While manifest cluster-mean centering is known to introduce Nickell's bias, we also showed how some implementations of the latent mean centering approach in the Bayesian framework may show bias, especially when the number of the time points is small. This is mostly due to problematic treatment of the unobserved presample outcome. The no centering approach that properly deals with the endogeneity problem (i.e., the *NC-ENDO model*) and the latent centering approach that does not utilize an auxiliary parameter for the unobserved presample outcome (i.e., the *ML-LC-ENDO-1 model*) performed best.

There are several limitations to this conclusion. First, we only considered a simple version of the ML-AR(1) model, with a fixed autoregressive parameter and a constant residual variance for example. Although it is likely that our findings still hold in the ML-AR(1) model which relaxes those restrictions, the implementation of the *NC-ENDO* or *ML-LC-ENDO-1 model* for such more complex setting in the traditional SEM-framework requires further investigation. It should be noted though that fitting those more complex models may be too demanding for the limited amount of information that is available when the number of time points is small. Second, we assumed that all subjects were measured simultaneously at equidistant time points. However, this may not always be true in practice: some time points may be scheduled further apart from each other or time points may differ between subjects. Treating the autoregressive parameter as time-specific adds further complications. Third, the ML-AR(1) model discussed in this paper did not include any other predictors in the model, either time-independent or time-varying. Clearly, our findings have implications

on more complex models such as the random intercept cross-lagged panel model (RI-CLPM) proposed by Hamaker et al. (2015), or dynamic network models (Bringmann et al., 2013), or autoregressive latent trajectory model (ALT) (Bollen & Curran, 2004). In case the number of time points is small, the bias in the autoregressive parameter may be substantial when the unobserved presample response and the endogeneity are not appropriately dealt with. Further research is required to explore the impact on the parameter estimates of other predictors. For example, it is possible that the cross-lagged effects will also show bias in the RI-CLPM (Allison et al., 2017).

Finally, it is worth noting that the ML-AR(1) is closely related to latent state-trait (LST) models (Steyer, Ferring, & Schmitt, 1992). For a recent discussion on the link between both frameworks, see for example Usami, Murayama, and Hamaker (2019). The LST framework can be used to study longitudinal dynamics of psychological attributes too and can, for example, determine the degree to which such attributes reflect stable effects, effects of person-situation interactions, or random measurement error. The latter two are not separated in the ML-AR(1) that we considered here. More knowledge may be gained on how reliable and stable estimation of the ML-AR(1) model is best achieved by comparing both frameworks (Lüdtke, Robitzsch, & Wagner, 2018).

References

Albert, T. (2009). *Bayesian computation with r* (2nd ed.). New York, NY: Springer Science & Business Media.

Allison, P., Williams, R., & Moral-Benito, E. (2017). Maximum likelihood for cross-lagged panel models with fixed effects. *Socius: Sociological Research for a Dynamic World*, *3*, 1-17. doi: 10.1177/2378023117710578

Asparouhov, T., Hamaker, E., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359-388. doi: 10.1080/10705511.2017.1406803

Asparouhov, T., & Muthén, B. (2010). *Bayesian analysis of latent variable models using mplus* (Tech. Rep.). Mplus Web Notes.

Baird, R., & Maxwell, S. (2016). Performance of time-varying predictors in multilevel models under an assumption of fixed or random effects. *Psychological Methods*, *12*(2), 175-188. doi: 10.1037/met0000070

Beaujean, A. (2014). *Latent variable modeling using r* (1st ed.). New York, NY: Routledge.

Bentler, P. (2004). *Eqs 6 structural equations program manual* (6th ed.). Encino, CA: Multivariate Software, Inc.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... Fox, J. (2011). Openmx: An open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306-317. doi: 10.1007/s11336-010-9200-6

Bollen, K., & Curran, P. (2004). Autoregressive latent trajectory (alt) models: A synthesis of two traditions. *Sociological methods and research*, *32*(3), 336-383. doi: 10.1177/0049124103260222

Bringmann, L., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., ... Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE*, *8*(4), e60188. doi: 10.1371/journal.pone.0060188

Flamant, N., & Soenens, B. (n.d.). *Consequences of coping with controlling teachers.*

Hallquist, M., & Wiley, J. (2018). Mplusautomation: An r package for facilitating large-scale latent variable analysis in mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 621-638. doi: 10.1080/10705511.2017.1402334

Hamaker, E., & Grasman, R. (2015). To center or not to center? investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*, 1492. doi: 10.3389/fpsyg.2014.01492

Hamaker, E., Kuiper, R., & Grasman, R. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102-116. doi: 10.1037/a0038889

Hamaker, E., & Muthén, B. (2019). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*.

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. (2015). A multilevel ar(1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, *50*(3), 334-349. doi: 10.1080/00273171.2014.1003772

Jöreskog, K., & Sörbom, D. (1996). *Lisrel 8: User's reference guide* (2nd ed.). Lincolnwood, IL: Scientific Software International, Inc.

Kenny, D., & Zautra, A. (2001). Trait-state models for longitudinal data [Proceedings Paper]. In Collins, LM and Sayer, AG (Ed.), *NEW METHODS FOR THE ANALYSIS OF CHANGE* (p. 243-263). 750 FIRST STREET NE, WASHINGTON, DC 20002 USA: AMER PSYCHOLOGICAL ASSOC. (Conference on New Methods for the Analysis of Change, PENN STATE UNIV, PHILADELPHIA, PA, 1998) doi: 10.1037/10409-008

Loeys, T., Josephy, H., & Dewitte, M. (2018). More precise estimation of lower-level interaction effects in multilevel models. *Multivariate Behavioral Research*, *53*(3), 335-347. doi: 10.1080/00273171.2018.1444975

Lüdtke, O., Robitzsch, A., & Wagner, J. (2018). More stable estimation of the starts model: A bayesian approach using markov chain monte carlo techniques. *Psychological Methods*, *23*, 570-593. doi: 10.1037/met0000155

McNeish, D. (2016). On using bayesian methods to address small sample problems.

Structural Equation Modeling: A Multidisciplinary Journal. doi: 10.1080/10705511.2016.1186549

McNeish, D. (2019). Two-level dynamic structural equation models with small samples. *Structural Equation Modeling: A Multidisciplinary Journal*. doi: 10.1080/10705511.2019.1578657

McNeish, D., & Stapleton, L. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*(4), 495-518. doi: 10.1080/00273171.2016.1167008

Muthén, B. (2010). *Bayesian analysis in mplus: A brief introduction* (Tech. Rep.). Mplus Web Notes.

Muthén, L., & Muthén, B. (2012). *Mplus user's guide: Statistical analysis with latent variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Exonometrica*, *49*(6), 1417-1426. doi: 10.2307/1911408

Plummer, M. (2015). Jags version 4.0.0 user manual [Computer software manual].

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, *69*(2), 167-190. doi: 10.1007/BF02295939

Raudenbush, S. (2004). *Hlm6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36. Retrieved from `http://www.jstatsoft.org/v48/i02/` doi: 10.18637/jss.v048.i02

SAS Institute, I. (2008). *Sas/stat 9.2 user's guide: The mixed procedure*. Cary,NC: SAS Institute Inc.

SAS Institute, I. (2013). *Sas/stat 13.1 user's guide: The calis procedure*. Cary, NC: SAS Institute Inc.

Schultzberg, M., & Muthén, B. (2018). Number of subjects and time points needed for multilevel time-series analysis: A simulation study of dynamic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4),

495-515. doi: 10.1080/10705511.2017.1392862

Steyer, R., Ferring, D., & Schmitt, M. (1992). States and treits in psychological assessment. *European Journal of Psychological Assessment*, *8*(2), 79-98.

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A Unified Framework of Longitudinal Models to Examine Reciprocal Relations [Article]. *PSYCHOLOGICAL METHODS*, *24*(5), 637-657. doi: 10.1037/met0000210

Zhang, Z., & Nesselroade, J. (2008). Bayesian estimation of categorical dynamic factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *42*(4), 729-756. doi: 10.1080/00273170701715998

Zitzmann, S., Lüdke, O., & Robitzsch, A. (2015). A bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research*, *50*(6), 688-705. doi: 10.1080/00273171.2015.1090899

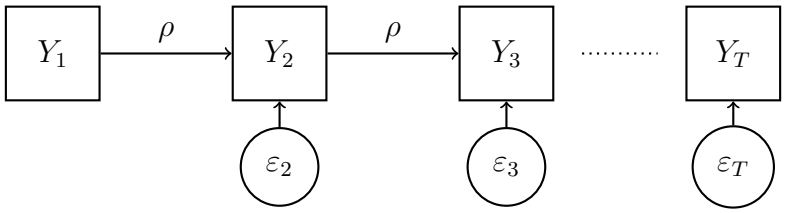
Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. (2016). A bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 661-679. doi: 10.1080/10705511.2016.1207179

Table 1

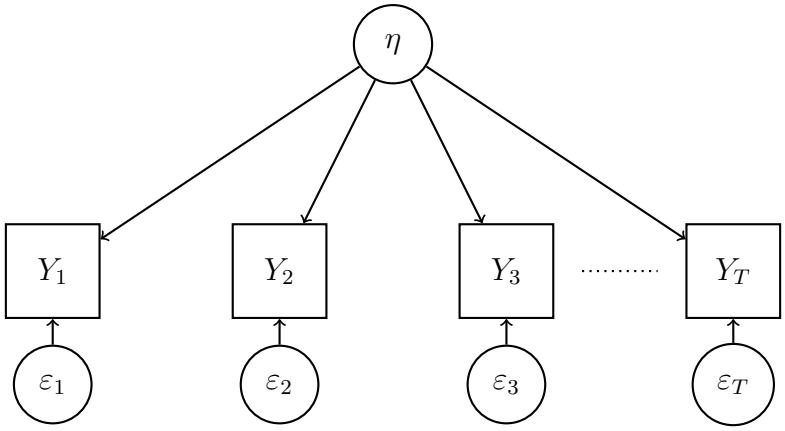
Estimates (and standard errors) of the parameters from the ML-AR(1) model on the empirical example with four time points using the different parametrizations.

	No centering approach			Centering approach	
	NC-EXO	NC-ENDO		CMC	DSEM
α	0.800(0.080)	2.376(0.243)	μ	3.360(0.052)	3.381(0.053)
ρ	0.760(0.023)	0.292(0.071)	ρ	-0.166(0.041)	0.747(0.026)
τ_α^2	<0.001(<0.001)	0.335(0.092)	τ_η^2	0.732(0.067)	0.120(0.074)
σ_ε^2	0.466(0.022)	0.343(0.027)	σ_ε^2	0.289(0.016)	0.459(0.023)

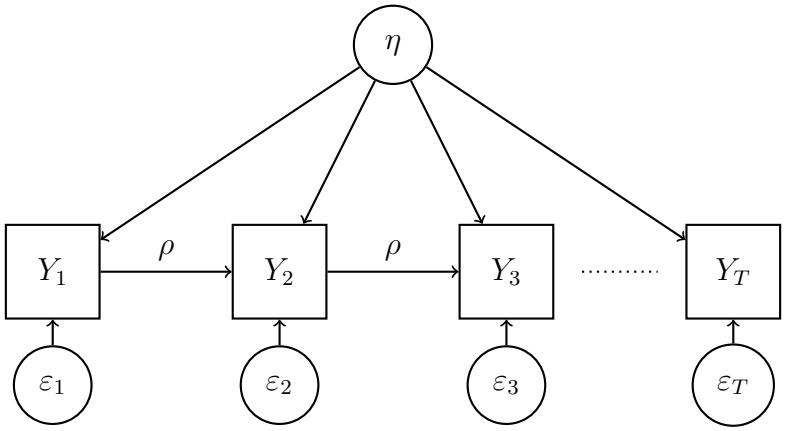
Notes. $\alpha = (1-\rho)\mu$ and $\tau_\alpha^2 = (1-\rho)^2\tau_\eta^2$



(a) Autoregressive process modeling state dependency



(b) The random intercept model of unobserved heterogeneity



(c) The combined state-dependency and unobserved heterogeneity model

Figure 1. The multilevel autoregressive model: a combination of state-dependency and trait

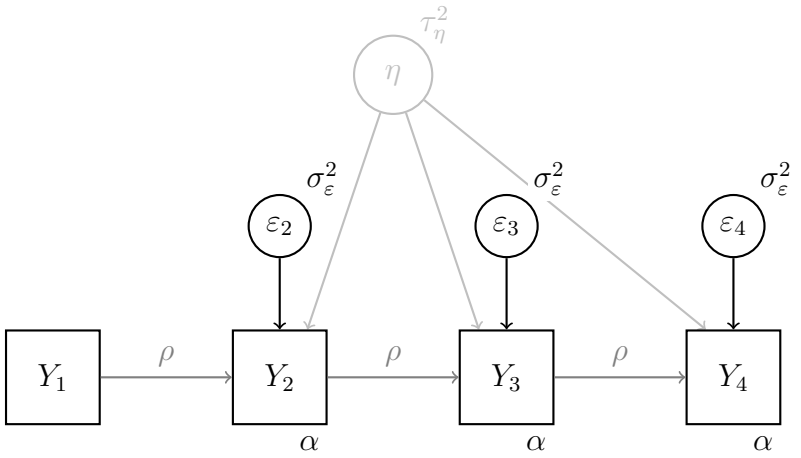


Figure 2. A schematic representation of the *NC-EXO* parametrization of the uncentered ML-AR(1) model for four time points.

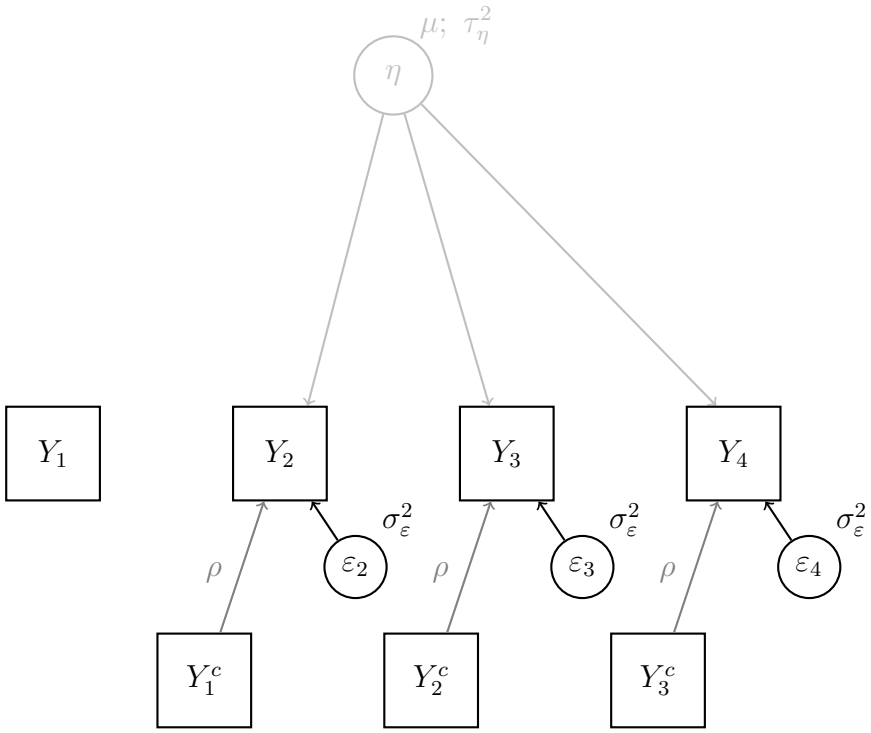


Figure 3. A schematic representation of the *CMC* parametrization of the cluster-mean centered ML-AR(1) model for four time points.

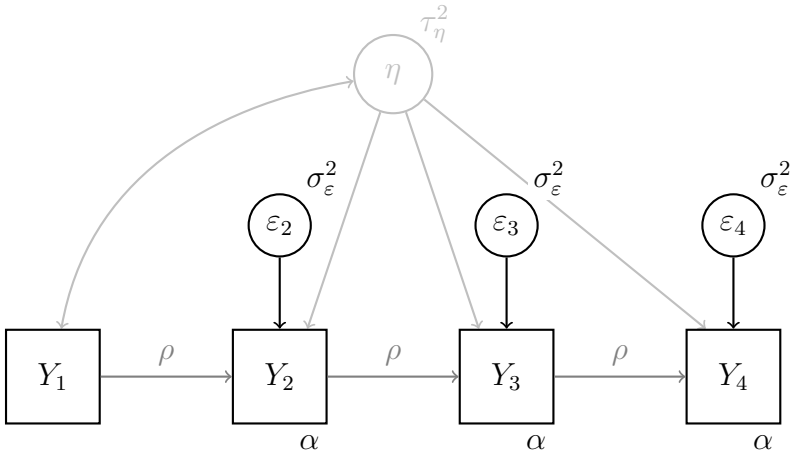


Figure 4. A schematic representation of the *NC-ENDO* parametrization of the uncentered ML-AR(1) model for four time points.

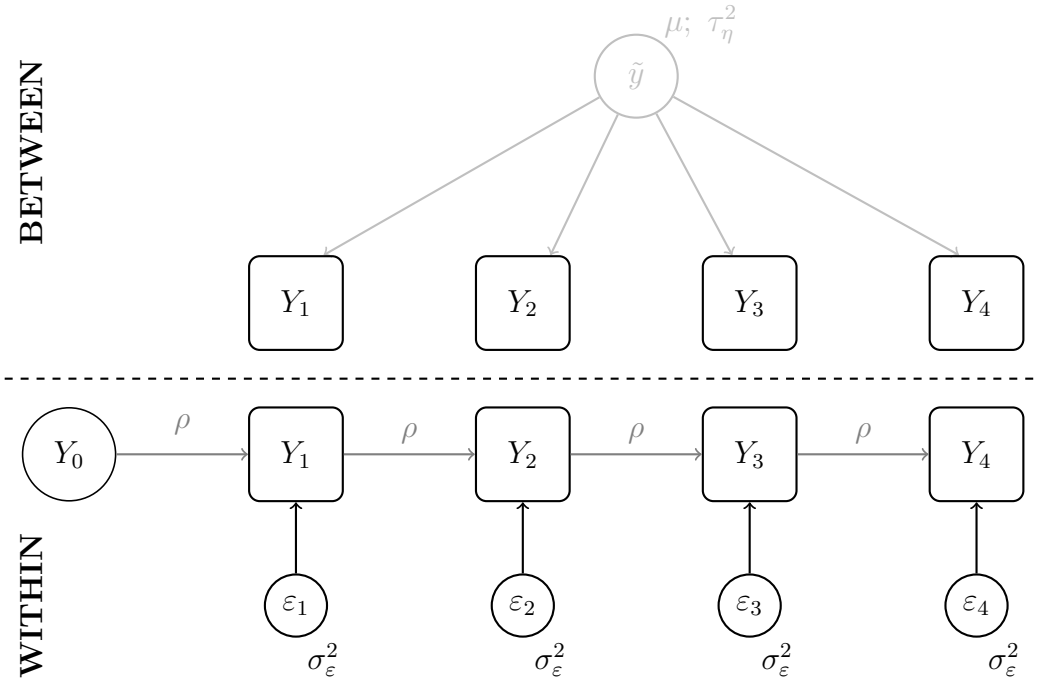


Figure 5. The *LC-ENDO model or DSEM*

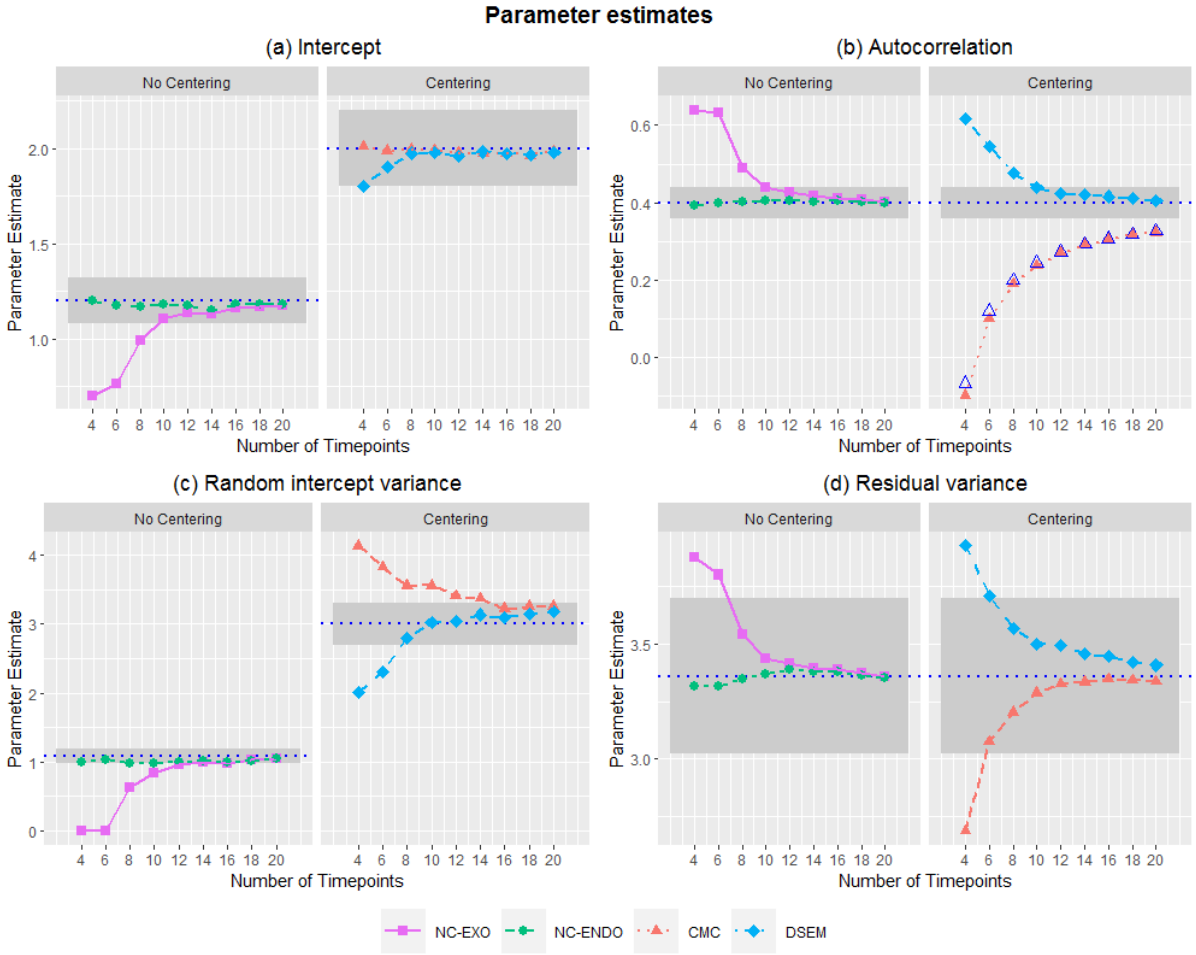
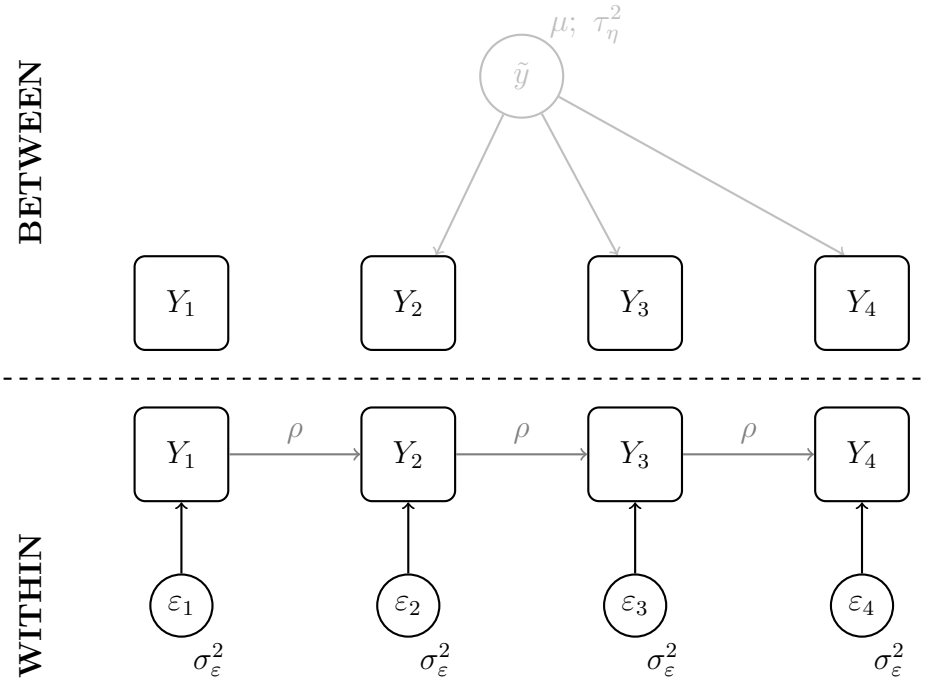
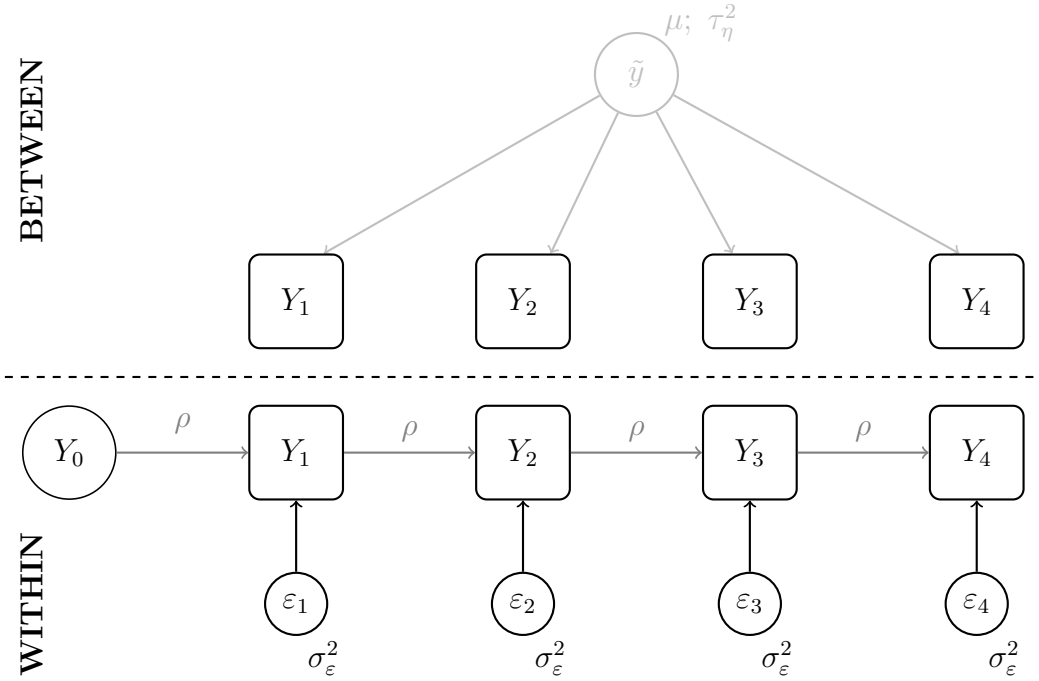


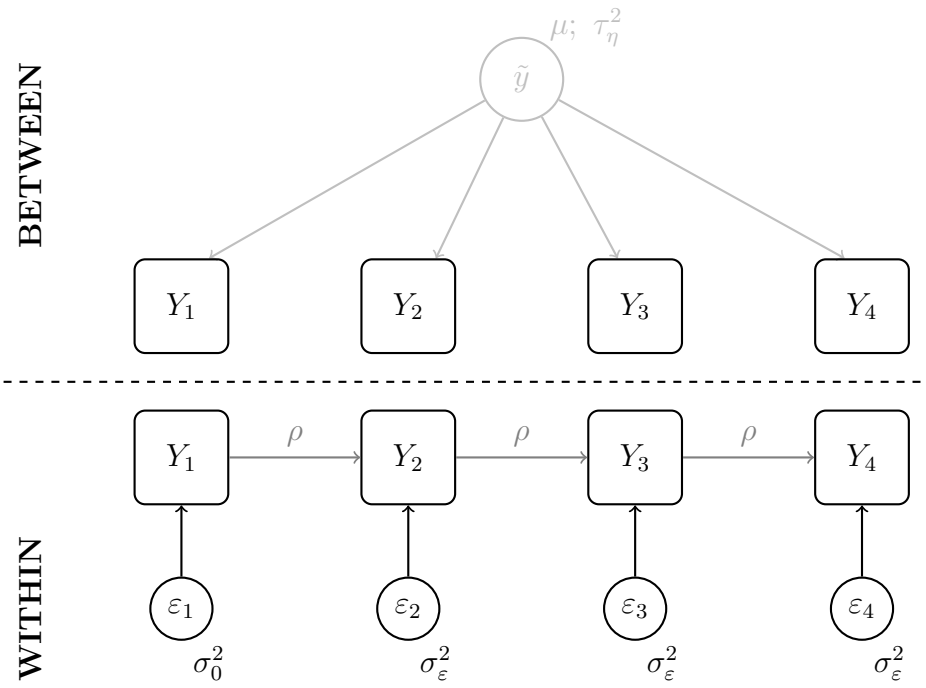
Figure 6. The parameter estimates of the ML-AR(1) model over 200 repetitions for $t = 4, 6, \ldots, 18, 20$ time points based on the different model parametrizations. The blue dotted line represents the true (transformed) model parameter. The gray band represents the absence of relative bias at a 10% cut-off. The blue triangles in subfigure (b) represents Nickell's bias.



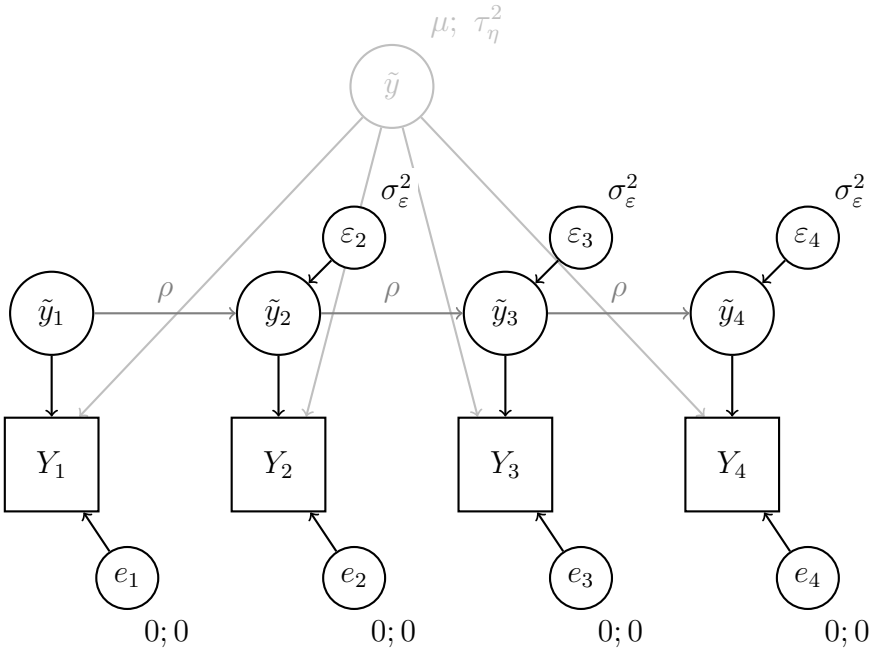
(a) The *BAY-LC-EXO model*



(b) The *BAY-LC-ENDO-0 model*



(c) The *BAY-LC-ENDO-1 model*



(d) The *ML-LC-ENDO-1 model*

Figure 7. A schematic representation of the different *LC models* of the ML-AR(1) model for four time points.

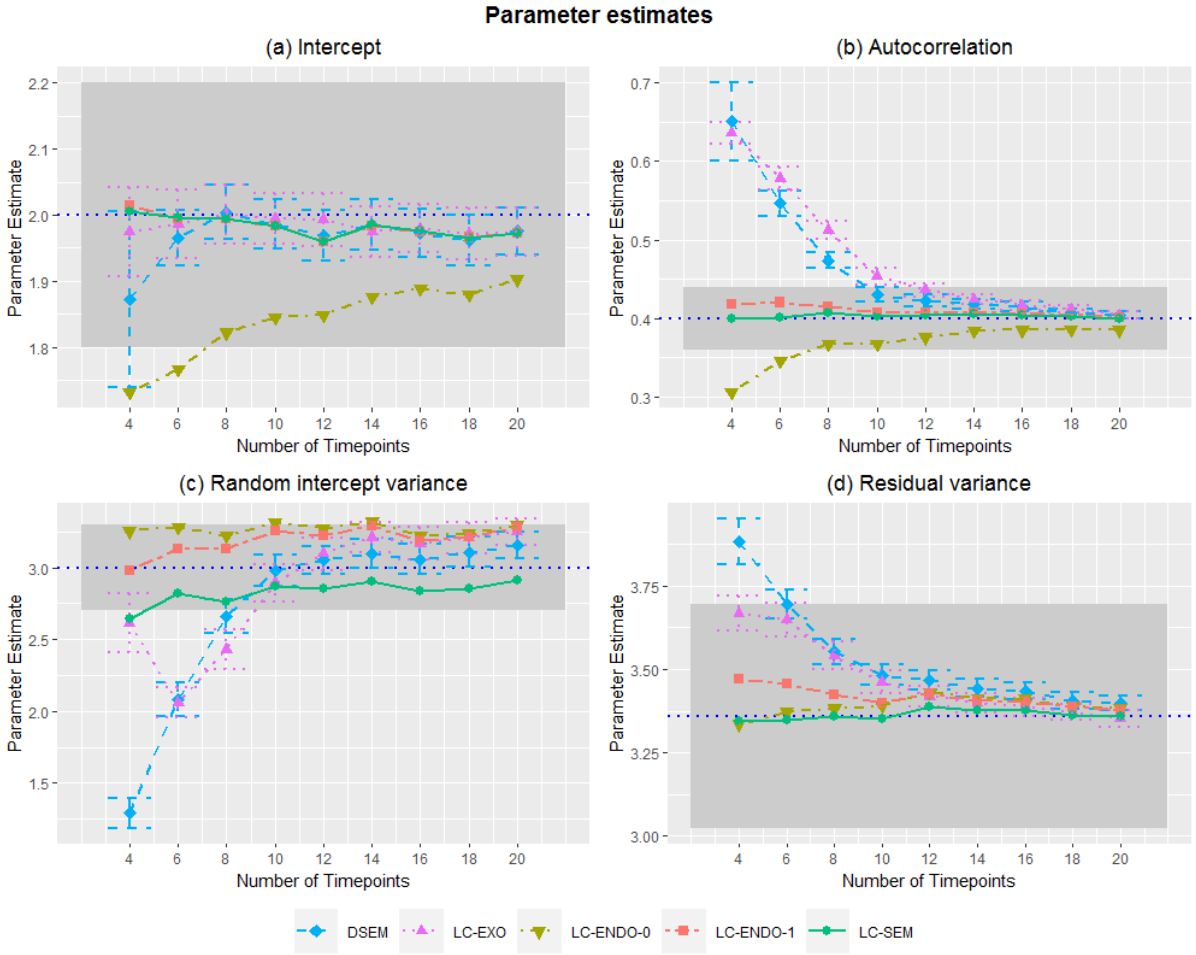


Figure 8. The parameter estimates of the ML-AR(1) model over 200 repetitions for $t = 4, 6, \ldots, 18, 20$ time points based on the different model implementations for the *LC model*. The blue dotted line represents the true model parameter. The gray band represents the absence of relative bias at a 10% cut-off. The 95%-confidence intervals for the DSEM and *BAY-LC-EXO model* has been provided as well.